Bilal **Chughtai**

brchughtaii@gmail.com | bilal-chughtai | bilalchughtai

Education

Trinity Hall, University of Cambridge — *Cambridge, UK*

BA & MMath in Mathematics — *2017 - 2021*

- Part IA - Class 1 (72%, unranked)
- Part IB - Class 1 (83%, ranking 17 of 240)
- Part II - Class 1 (unranked due to COVID-19 pandemic)
- Part III - Distinction (88%, ranking 22 of 272, graduated aged 20)
  Examinable courses: Quantum Field Theory, General Relativity, Cosmology, Field Theory in Cosmology.
  Essay: "The Search for CMB B-Mode Polarization from Inflationary Gravitational Waves" (90%)

Saint Olave's Grammar School — *London, UK*

Secondary and Sixth Form Education — *2010 - 2017*

A levels:	Mathematics A* • Further Mathematics A* • Physics A* • Chemistry A* • Biology AS A • EPQ A
University Entrance Exams:	STEP1 S (117/120) • STEP2 S (116/120) • STEP3 1 (92/120) • MAT 74/100
GCSEs:	7 A*s • 4 As • FSMQ A

Publications

[1] **Bilal Chughtai**, Alan Cooney and Neel Nanda. Summing Up The Facts: Additive Mechanisms behind Factual Recall. *NeurIPS 2023 Attributing Model Behaviour at Scale Workshop.*

[2] Dane Sherburn, **Bilal Chughtai** and Owain Evans. Language Models Struggle to Explain Themselves. *Preprint. Under Review.*

[3] **Bilal Chughtai**, Lawrence Chan and Neel Nanda. A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations. *ICML 2023. ICLR 2023 Physics4ML Workshop (spotlight).* arXiv:2302.03025.

Experience

AI Safety Researcher — *London, UK*

London Initiative for Safe AI, Grant Funded — *March 2023 - Present*

- Led mechanistic interpretability project investigating recall of memorised facts in transformer based language models. We find several independent circuits that interact *additively* – each circuit is insufficient alone, but their sum suffices to solve the task as a result of constructive interference. We term this general phenomena the 'additive motif'. Proposed the project, ran mechanistic interpretability experiments using TransformerLens and PyTorch, and wrote the large majority of the paper.
- Contributor to evaluations paper investigating the capability of modern language models to honestly explain their reasoning processes. We create a benchmark evaluating the faithfulness of model written explanations on a dataset of text-based binary classification tasks. We find modern LLMs struggle to produce faithful explanations matching their IID and OOD behaviour. Ran experiments using the OpenAI API, clarified and distilled results, and wrote a large portion of the paper.
- Projects in progress include situational awareness evaluations, and various mechanistic interpretability projects.
- Mentoring junior mechanistic interpretability researchers in research projects.
- Reviewer for the Socially Responsible Language Modelling Research NeurIPS workshop.

Stanford Existential Risks Initiative — *London, UK*

ML Alignment Theory Scholar — *October 2022 - January 2023*

- Research scholar in Neel Nanda's mechanistic interpretability stream.
- Led mechanistic interpretability project investigating grokking in small models trained on group theoretic tasks. We build on prior work understanding grokking of modular addition. Modular addition can be thought of as composition in the cyclic group. We generalise this task and reverse engineer how networks implement composition in *arbitrary* groups, find a consistent representation theory based algorithm, and then use this understanding to investigate the universality hypothesis of mechanistic interpretability.

Met Office — *Exeter, UK*

Foundation (Junior) Scientific Software Engineer (SSE) — *January 2022 - June 2022*

- Member of the ExCALIBUR (Exascale Computing ALgorithms and Infrastrucures benefiting UK Research) pool of deployable SSEs.
- Worked in the LFRic project on infrastructure supporting next generation weather modelling systems, optimised for new supercomputer hardware. Development in Python and Fortran.

Metaswitch Networks — *London, UK*

Software Engineering Intern — *June 2019 - August 2019*

- Worked in the Support Tools team within Information Systems on enhancements to a Google Cloud based Business Intelligence platform.
- Migrated old C# Extract-Transfer-Load (ETL) processes uploading data to Google BigQuery to a newer, more robust infrastructure. Cleaned and used new data source to produce dynamic reports for internal end users.
- Wrote python scripts to enhance permission deployment to our Google Project, to both improve speed and to reflect organisational changes better through use of active directory groups over raw emails. Deployed as jobs on GitLab CI/CD pipeline and also to run hourly via a dockerised crontab job.

Private Tutor *London, UK*

Self Employed *May 2020 - May 2022*

- Over 200 hours of tutoring experience, mostly in GCSE and A level Maths. Wrote preparatory resources and tutored students in university entrance exams including the MAT, ENGAA and STEP. Provided mock interviews to university applicants in the physical sciences.

Cambridge Assessment *Cambridge, UK*

Assistant Examiner *July 2021, July 2022*

- Worked in several small teams marking in excess of 800 candidate scripts annually for the STEP Mathematics II and III exams.

Courses

Alignment Research Engineer Accelerator *London, UK*

Participant *May - June 2023*

- The Alignment Research Engineer Accelerator (ARENA) is a 6 week machine learning engineering bootcamp, focusing on AI Safety.
- Replicated several landmark mechanistic interpretability papers (e.g. induction heads, IOI, OthelloGPT) and reinforcement learning papers (e.g. PPO, RLHF). Trained models at scale across multiple GPUs.

Cambridge AI Safety Hub – CaMLAB *Cambridge, UK*

Teaching Assistant *March 2023*

- TA on course 'Cambridge Machine Learning for Alignment Bootcamp'. Course teaches ML fundamentals, and the basics of Mechanistic Interpretability and Reinforcement Learning.
- Gave talks introducing new conceptual content, and provided one to one mentorship to individuals participating in the bootcamp.

Center for AI Safety *Remote*

ML Safety Scholars Program *June 2022 - February 2023*

- Intense fully-funded 10 week summer program. Took online courses in the fundamentals of deep learning and ML Safety. Ranked in the top quartile in assignments.
- Implemented various DL architectures from scratch in PyTorch, including MLPs, CNNs, RNNs and Transformers.
- TA'd a later version of the course.

Volunteering

Cambridge University Powerlifting Club *Cambridge, UK*

Webmaster *May 2021 - Present*

- Designed and deployed a new club website (cuplc.co.uk), reducing club costs and increasing maintainability for future committees. A web development project built using Ruby, Jekyll, HTML and CSS.
- Automated previously manually curated club student and alumni records through open source OpenPowerlifting data. Updates daily.
- Organized logistics for the 2022 Varsity Powerlifting competition against Oxford, and helped run the event on the day.
- Assist in general running of the club, including running a freshers fair stall and introductory sessions, team training, in house competitions and socials.

Trinity Hall JCR *Trinity Hall, Cambridge, UK*

Webmaster *November 2017 - March 2020*

- Responsible for maintaining and updating the JCR website, as well as aiding the committee with technical tasks.
- Ran technical aspects of the yearly college room ballot, the process by which students choose their room for the following academic year. Worked closely with several different groups of people, balancing requirements. System consisted of a database storing the state of the ballot linked via a python script to a custom web app that displays this state in a more user friendly form, by overlaying live properties (availability/occupant, weekly rent, ensuite, etc.) of the rooms onto a floor plan.
- Sat on the college IT Advisory Group, representing the undergraduate body regarding IT matters.
- Helped organize and run freshers week events for new undergraduates at Trinity Hall.

Skills

Programming Python • MATLAB • Jekyll • LaTeX • HTML • CSS • SQL • Git

Machine Learning PyTorch • W&B

Honors & Awards

2024 **Research Grant**, 6 months of funding to visit David Bau's AI interpretability lab *LTFF*

2023 **Research Grant**, 6 months of funding for AI interpretability and evaluations projects *LTFF*

2021 **Parks Prize for Mathematics**, for "obtaining a particularly strong result in Tripos Examinations" *Cambridge, UK*

2020 **Wylie Prize for Mathematics**, for "obtaining a particularly strong result in Tripos Examinations" *Cambridge, UK*

2020, 2021 **Trinity Hall Bateman Scholarship**, for "obtaining a first class result in Tripos Examinations" *Cambridge, UK*

2018, 2019 **Trinity Hall Scholarship**, for "obtaining a first class result in Tripos Examinations" *Cambridge, UK*

2017 **HG Abel Prize**, for "outstanding A level results", awarded by Saint Olave's Grammar School *London, UK*

2017 **Highest Performer at A-level**, Saint Olave's Grammar School *London, UK*